09/27/00 JC715 U.S. PTO

S&H Form: PTO/SB/05 (9/99)

UTILITY PATENT APPLICATION TRANSMITTAL

(Only for new nonprovisional applications under 37 CFR 1.53(b))

Attorney Docket No. 826.1628/JDH	
First Named Inventor or Application Identifier: Jun IBUKI, et al.	
Express Mail Label No.	

JC711 U.S. PTO 09/669897 09/27/00

APPLICATION ELEMENTS
See MPEP chapter 600 concerning utility patent application contents.

ADDRESS TO: **Assistant Commissioner for Patents**
Box Patent Application
Washington, DC 20231

1. [X] Fee Transmittal Form
2. [X] Specification, Claims & Abstract ..... [Total Pages: 44]
3. [X] Drawing(s) *(35 USC 113)* ..... [Total Sheets: 20]
4. [ ] Oath or Declaration ..... [Total Pages: _]
   a. [ ] Newly executed (original or copy)
   b. [ ] Copy from a prior application (37 CFR 1.63(d)) *(for continuation/divisional with Box 17 completed)*
      i. [ ] DELETION OF INVENTOR(S)
      Signed statement attached deleting inventor(s) named in the prior application, see 37 CFR 1.63(d)(2) and 1.33(b).
5. [ ] Incorporation by Reference (usable if Box 4b is checked)
   The entire disclosure of the prior application, from which a copy of the oath or declaration is supplied under Box 4b, is considered as being part of the disclosure of the accompanying application and is hereby incorporated by reference therein.
6. [ ] Microfiche Computer Program *(Appendix)*
7. [ ] Nucleotide and/or Amino Acid Sequence Submission *(if applicable, all necessary)*
   a. [ ] Computer Readable Copy
   b. [ ] Paper Copy (identical to computer copy)
   c. [ ] Statement verifying identity of above copies

ACCOMPANYING APPLICATION PARTS

8. [ ] Assignment Papers (cover sheet & document(s))
9. [ ] 37 CFR 3.73(b) Statement *(when there is an assignee)* [ ] Power of Attorney
10. [ ] English Translation Document *(if applicable)*
11. [ ] Information Disclosure Statement (IDS)/PTO-1449[ ] Copies of IDS Citations
12. [ ] Preliminary Amendment
13. [X] Return Receipt Postcard (MPEP 503) *(Should be specifically itemized)*
14. [ ] Small Entity Statement(s) [ ] Statement filed in prior application, status still proper and desired.
15. [X] Certified Copy of Priority Document(s) *(if foreign priority is claimed)*
16. [ ] Other:

17. If a CONTINUING APPLICATION, *check appropriate box and supply the requisite information:*

[ ] Continuation [ ] Divisional [ ] Continuation-in-part (CIP) of prior application No: ____/_____

18. CORRESPONDENCE ADDRESS

21171
PATENT TRADEMARK OFFICE

Staas & Halsey

DC#260459v1

NEW APPLICATION FEE TRANSMITTAL		Attorney Docket No.	826.1628/JDH
		Application Number	To be assigned
		Filing Date	September 27, 2000
AMOUNT ENCLOSED	$768.00	First Named Inventor	Jun IBUKI, et al.

FEE CALCULATION (fees effective 12/29/99)

CLAIMS	(1) FOR	(2) NUMBER FILED	(3) NUMBER EXTRA	(4) RATE	(5) CALCULATIONS
	TOTAL CLAIMS	11 - 20 =	0	X $ 18.00 =	$ 0.00
	INDEPENDENT CLAIMS	4 - 3 =	1	X $ 78.00 =	78.00
	MULTIPLE DEPENDENT CLAIMS (any number; if applicable)			+ $260.00 =	
			BASIC FILING FEE		690.00
			Total of above Calculations =		$ 768.00
	Surcharge for late filing fee, Statement or Power of Attorney ($130.00)				+
	Reduction by 50% for filing by small entity (37 CFR 1.9, 1.27 & 1.28).				-
			TOTAL FILING FEE =		$ 768.00
	Surcharge for filing non-English language application ($130.00; 37 CFR 1.52(d))				+
	Recordation of Assignment ($40.00; 37 CFR 1.21(h)(1))				0.00
			TOTAL FEES DUE =		$ 768.00

METHOD OF PAYMENT

[X] Check enclosed as payment.

[ ] Charge "TOTAL FEES DUE" to the Deposit Account No., below.

[ ] No payment is enclosed and no charges to the Deposit Account are authorized at this time.

GENERAL AUTHORIZATION

[X] If the above-noted "AMOUNT ENCLOSED" is not correct, the Commissioner is hereby authorized to credit any overpayment or charge any additional fees necessary to:

Deposit Account No.	19-3935
Deposit Account Name	STAAS & HALSEY LLP

[X] The Commissioner is also authorized to credit any overpayments or charge any additional fees required under 37 CFR 1.16 (filing fees) or 37 CFR 1.17 (processing fees) during the prosecution of this application, including any related application(s) claiming benefit hereof pursuant to 35 USC § 120 (e.g., continuations/divisionals/CIPs under 37 CFR 1.53(b) and/or continuations/divisionals/CPAs under 37 CFR 1.53(d)) to maintain pendency hereof or of any such related application.

SUBMITTED BY: STAAS & HALSEY LLP

Typed Name	James D. Halsey, Jr.	Reg. No.	22,729
Signature		Date	September 27, 2000

© 2000 Staas & Halsey

0024250 "768.00 960

APPLICATION FOR

UNITED STATES LETTERS PATENT

SPECIFICATION

09659960 "45897 09250 00

Inventor(s): Jun IBUKI, Ryou OCHITANI and
Fumihito NISINO

Title of the Invention: FACT DATA UNIFYING METHOD AND
APPARATUS

FACT DATA UNIFYING METHOD AND APPARATUS

Background of the Invention

Field of the Invention

The present invention relates to a fact data unifying method and apparatus which extracts a description of a fact within a document, puts the extracted description into a database as a set of data having consistency, and detects or corrects a corresponding error included in an original text based on an inconsistent point of fact data.

Description of the Related Art

A variety of methods were conventionally proposed as a technique extracting information within a text. By way of example, for data in compliance with a predetermined framework such as new product information, organism information, etc., a correspondence table between an expression format and data to be extracted within a text is stored, and corresponding data is extracted when a match is found for the expression format stipulated by scanning a text.

Assume that a correspondence table shown in Fig. 1A is stored, and fact data which is composed of a target object, an attribute name, and an attribute value, and

09699897.092700

is shown in Figs. 1B and 1C, is extracted. In this example, "a new president of a company C" and "a person D is assigned" respectively match *1 and *2 in the correspondence table. Therefore, "company C" is extracted as a target object, "representative" is extracted as an attribute name, and "person D" is extracted as an attribute value.

If a target is limited to an error on a representation level included in a text, various error correction techniques already exist. By way of example, a method registering an expression included in a text, and pointing to an unregistered word, a method pointing to representation fluctuations, etc. are known.

As described above, fact data extraction from a text is widely used. However, it is not always possible to obtain information desired to view only from the information from one point within a text. Therefore, data from the whole of a text must normally be unified.

Generally, however, data to be extracted includes a considerable number of errors (or data inconsistencies) such as an error included in a text itself, an error in an extraction process, etc., (or data inconsistency). Since errors must manually be checked and removed, or rewritten, data cannot simply be aggregated.

0965997.090200

Summary of the Invention

The present invention was developed in consideration of the above described background, and aims at enabling suitable data to be aggregated by correcting or standardizing an error or representation fluctuations within extracted data due to an incorrect description within a text or an error in an extraction process.

Fig. 2 is a block diagram showing the fundamental configuration of the present invention. In this figure, 1 is a data extracting unit extracting from a text fact data stipulated by a combination of three such as a target object, an attribute name, and an attribute value; 2 is a data aggregating unit grouping the same data, and aggregating the number of occurrences; 3 is an inconsistency detecting unit detecting a group of inconsistent data that is inconsistent as a result of scanning the data set concerning the same object within the output of the data aggregating unit; 4 is a correctness/incorrectness determining unit determining which data is correct within an inconsistent data group; 5 is an integrating unit integrating data aggregated by the data aggregating unit, and data determined to be correct by the

correctness/incorrectness determining unit.

Furthermore, 6 is a reliability degree assigning unit assigning the degree of reliability to fact data when the fact data is extracted from a text; 7 is a data unifying unit unifying similar data into one data; 8 is an error pattern removing unit discarding as an error fact data which matches a pre-registered error pattern; and 9 is a determination method deciding unit deciding a correctness/incorrectness determination method executed by the correctness/incorrectness determining unit.

As shown in Fig. 2, according to the present invention, the above described problems are solved as follows.

(1) The data extracting unit 1 extracting from a text fact data stipulated by a combination of three such as a target object, an attribute name, and an attribute value; the data aggregating unit 2 grouping the same data throughout a text, and aggregating the number of occurrences; the inconsistency detecting unit 3 detecting an inconsistent data group by scanning a data set concerning the same object within the output of the data aggregating unit 2, and; the correctness/incorrectness determining unit 4 determining which data is correct within the

inconsistent data group detected by the inconsistency detecting unit 3; and the final data integrating unit 5 integrating the correct data aggregated by the data aggregating unit 2, and the data determined to be correct by the correctness/incorrectness determining unit 4 are comprised, so that suitable data can be unified by removing error data from extracted fact data.

(2) In the above provided (1), the reliability degree assigning unit 6 assigning the degree of reliability to fact data when the fact data is extracted from a text is further comprised. When the number of occurrences is aggregated by the data aggregating unit 2, the degree of reliability of aggregated data is calculated from the degrees of reliability of individual data, and the calculated degree of reliability is assigned to an aggregation result. The correctness/incorrectness determining unit 4 determines whether each data within a data group is either correct or incorrect by using the degree of reliability assigned to the aggregated data, leading to an improvement in the correctness/incorrectness determination.

(3) In the above provided (2), the reliability degree assigning unit 6 is configured by an event type extracting unit determining the type of event information possessed by a text, which is determined

0042507.0965960

to be an extraction target when fact data is extracted from the text, and a reliability degree evaluating unit evaluating the degree of reliability from an event type based on a correspondence table between an event type and the degree of reliability, so that an accurate degree of reliability is assigned.

(4) In the above provided (2), the reliability degree assigning unit 6 is configured by an attention degree evaluating unit calculating the degree of attention to a target object being an extraction target within a text, and a reliability degree evaluating unit evaluating the degree of reliability of data based on the degree of attention, so that an accurate degree of reliability is assigned.

(5) In the above provided (2), the reliability degree assigning unit 6 is configured by a correspondence table between the bibliographical information such as an issuance source, an author, etc. of a text, and the degree of reliability of each data described in the text, and a reliability degree evaluating unit evaluating the degree of reliability of a text based on the bibliographical information of the text by referencing the correspondence table between the bibliographical information and the degree of reliability, so that the degree of reliability for which a general tendency is

0969907-0020-00

considered based on an author, an issuance source, etc. is assigned.

(6) In the above provided (5), a correctness/incorrectness flag is attached to the fact data extracted by the data extracting unit 1, the fact data attached with the correctness/incorrectness flag is input, and an expectation value of correctness/incorrectness of data having a particular attribute value is calculated for each attribute name of fact data, and a correspondence table between bibliographical information and the degree of reliability is generated, so that a correspondence table between an attribute value and the degree of reliability is semi-automatically generated from a text.

(7) In the above provided (1) through (6), an attribute/determination method correspondence table making a correspondence between a target object, an attribute name, and a determination method used when a correctness/incorrectness determination is made; and a determination method deciding unit deciding a correctness/incorrectness determination method according to an attribute name based on the attribute/determination method correspondence table are arranged. The correctness/incorrectness determining unit makes a correctness/incorrectness

0099769.0002.00

determination with the determination method specified by the determination method deciding unit when an inconsistent data group is input, so that a flexible correctness/incorrectness determination according to an attribute is made.

(8) In the above provided (1) through (7), an error pattern removing unit is arranged between the data extracting unit 1 and the inconsistency detecting unit 3. The error pattern removing unit 8 determines whether each data is either correct or incorrect by making a matching between the fact data extracted by the data extracting unit 1 and a pre-registered error pattern. If the extracted fact data matches a pre-registered error pattern, the data is determined to be incorrect and is discarded, and only the data determined to be correct is transmitted to the inconsistency detecting unit 3, whereby an error that the error removing unit can determine alone is removed.

(9) In the above provided (1) through (6), the data unifying unit 7 is arranged after the data aggregating unit 2. The data unifying unit 7 unifies similar data into one data, and passes the unified data to the inconsistency detecting unit 3, so that fluctuations caused by different expressions of the same object are absorbed.

0966997*0927000

Brief Description of the Drawings

Figs. 1A through 1D explain the method extracting information within text;

Fig. 2 is a block diagram showing the fundamental configuration of the present invention;

Fig. 3 exemplifies the configuration of a system performing a fact data unifying process;

Fig. 4 shows a first preferred embodiment according to the present invention;

Figs. 5A through 5E exemplify the process performed in the first preferred embodiment;

Fig. 6 is a flowchart showing the process performed in the first preferred embodiment;

Fig. 7 is a block diagram showing the functions of a second preferred embodiment according to the present invention;

Fig. 8 exemplifies a first internal configuration of a reliability degree assigning unit;

Figs. 9A through 9D exemplify a process performed by the reliability degree assigning unit shown in Fig. 8 (No. 1);

Figs. 10A and 10B exemplify a process performed by the reliability degree assigning unit shown in Fig. 8 (No. 2);

09999997.092700

Fig. 11 exemplifies a second internal configuration of the reliability degree assigning unit;

Figs. 12A through 12D exemplify a process performed by the reliability degree assigning unit shown in Fig. 11;

Fig. 13 exemplifies a third internal configuration of the reliability degree assigning unit;

Figs. 14A through 14F exemplify a process performed by the reliability degree assigning unit shown in Fig. 13;

Fig. 15 exemplifies the configuration for generating a correspondence table between bibliographic information and the degree of reliability;

Fig. 16 exemplifies a third preferred embodiment according to the present invention;

Fig. 17 exemplifies a fourth preferred embodiment according to the present invention;

Figs. 18A through 18C exemplify an error pattern determination in the fourth preferred embodiment;

Fig. 19 shows a fifth preferred embodiment according to the present invention; and

Figs. 20A through 20C exemplify a process performed in the fifth preferred embodiment.

0026507.0026950

Description of the Preferred Embodiments

Hereinafter, preferred embodiments according to the present invention will be explained.

Fig. 3 exemplifies the configuration of a system performing a fact data unifying process, according to the present invention. In this figure, 101 is an input/output device composed of a display device such as a CRT, a liquid crystal display, etc., and an input device for inputting characters, symbols, commands, etc., such as a keyboard, a mouse, etc.; 102 is a CPU; 103 is a memory composed of a ROM, a RAM, etc.; 104 is an external storage device storing programs, data, etc.; 105 is a medium reading device reading/writing data by accessing a portable storage medium such as a floppy disk, an MO, a CD-ROM, etc.; and 106 is a communications interface including a modem making a data communication by using a telephone line, a network card for making a data communication by using a network such as a LAN, etc.

The external storage device 104 stores the programs performing a fact data unifying process according to the present invention, text data from which fact data is extracted, unified data obtained as a result of performing the fact data unifying process, and the like.

09/260,768 03/02/99

Fig. 4 is a block diagram showing the functions of a first preferred embodiment according to the present invention. The first preferred embodiment is explained with reference to this figure.

In Fig. 4, 11 is a data extracting unit analyzing a description of fact data within a text, and extracting the description as fact data; 12 is a data aggregating unit grouping data of the same type among the fact data extracted by the data extracting unit 11 into one data, and counting the number of occurrences of each fact data; 13 is an inconsistency detecting unit searching for an inconsistency (such as a combination of inconsistent fact data which cannot be consistent) within a set of fact data extracted from a text; 14 is a correctness/incorrectness determining unit determining which inconsistent data detected by the inconsistency detecting unit 13 is correct/incorrect; and 15 is a final data integrating unit integrating and presenting data determined to be correct.

In Fig. 4, when text data is input, the data extracting unit 12 analyzes a description within the text, and extracts the description as fact data, similar to the method explained in the above described conventional example.

Fig. 5A is an output of the data extracting unit

0025700-0969769

12 in the case where the correspondence table shown in Fig. 1A is used, and fact data in an expression format stipulated in the correspondence table is extracted from a text. According to this correspondence table, fact data composed of target objects (company A, company F, ……, company H), attributes names (representative,……, location), and attribute values (country B, country G, …… country C) is extracted as shown in Fig. 5A.

The data aggregating unit 12 sorts the fact data, groups the same data, and counts the occurrences of each fact data. Fig. 5B exemplifies an output of the data aggregating unit 12 with regard to the fact data shown in Fig. 5A. As shown in this figure, target objects, attribute names, attribute values, and the numbers of occurrences of the fact data matching the target object, attribute names, and the attribute values are output.

The inconsistency detecting unit 13 detects inconsistent data within a fact data set. For this detection, by way of example, the following process is performed.

i) The following operations are repeated for all of target objects within a data set.

ii) The following operations are repeated for all of attribute names possessed by selected target objects.

iii) If there are a plurality of attribute

0995997-090200

values corresponding to the same attribute name, the corresponding data group is output as an inconsistent data group, and others are output as consistent data.

Fig. 5C exemplifies inconsistent data detected by the inconsistency detecting unit 13. As shown in this figure, there are two attribute value types "B" and "D" for the target object "company A" and an attribute name "representative" among the fact data aggregated by the data aggregating unit 12 as shown in Fig. 5C. Therefore, the attribute values "B" and "D" are detected as inconsistent data, and transmitted to the correctness/incorrectness determining unit 14. The rest of the data aggregated by the data aggregating unit 12 is transmitted to the final data integrating unit 15 as consistent data.

The correctness/incorrectness determining unit 14 determines which inconsistent data is correct/incorrect.

For this process, the following diversified algorithms are considered.

i) Data having the maximum number of occurrences within a group is determined to be correct, and the others are determined to be incorrect.

ii) Data having the number of occurrences, which is equal to or larger than a particular threshold value, is

determined to be correct, and the others are determined to be incorrect.

Fig. 5D exemplifies an output of the correctness/incorrectness determining unit 14. This is an example of an output in the case where a correctness/incorrectness determination is made with the algorithm provided in the above described i).

The number of occurrences of the attribute value "B" is 2, and that of "D" is 1 among the attribute values "B" and "D" of the target object "company A" and the attribute name "representative", which are detected as inconsistent data. Therefore, in this example, the attribute value "B" is adopted as "correctness", whereas the attribute value "D" is discarded as "incorrectness" as shown in Fig. 5D.

The final data integrating unit 15 integrates and presents the data transmitted from the inconsistency detecting unit 13 as consistent data, and the data determined to be correct by the correctness/incorrectness determining unit 14. Fig. 5E exemplifies an output of the final data unifying unit 15. As shown in this figure, data which is transmitted from the inconsistency detecting unit 13 as consistent data, and data which is determined to be correct by the correctness/incorrectness determining unit 14 among

0999997.0992001

the data aggregated by the data aggregating unit 12 are output as correct data.

Fig. 6 is a flowchart showing the process performed in this preferred embodiment. This process is explained with reference to this figure.

In Fig. 6, a description of fact data within input text data is analyzed and extracted as fact data in step S1, so that, for example, the fact data shown in Fig. 5A are obtained.

In step S2, extracted fact data are sorted according to target objects, attribute names, and attribute values, and the numbers of occurrences of the sorted data are counted. As a result, the data shown in Fig. 5B is obtained.

In step S3, one of the sorted target objects is extracted. In step S4, one of attribute names for the extracted target object is selected. In step S5, its consistency is checked. For example, if the inconsistent data shown in Fig. 5C is detected, the process goes to step S6 where it is determined whether the inconsistent data is either correct or incorrect with the algorithm provided in the above described i) or ii), and incorrect data is discarded. If the data is determined to be consistent, this data is integrated in step S7.

In step S8, it is determined whether or not the

0096598.0026700

consistency checking is completed for the attribute names. If the consistency checking is not completed, the process goes back to step S4, and the above described operations are repeated. If the consistency checking for the attribute names is determined to be completed, it is determined whether or not the consistency checking for the target objects is completed in step S9. If the consistency checking is not completed, the process goes back to step S3 and the above described operations are repeated. If the consistency checking is determined to be completed for the target objects, the process is terminated.

Fig. 7 is a block diagram showing the functions of a second preferred embodiment according to the present invention. This preferred embodiment is implemented by adding a reliability degree assigning unit to the first preferred embodiment so as to assign the degree of reliability of text data, and is intended to make a correctness/incorrectness determination based on the degree of reliability.

In this figure, the data extracting unit 11 analyzes a description of fact data within a text, and extracts the description as fact data as described above. Additionally, a reliability degree assigning unit 16 evaluates the degree of reliability of extracted data

0956997 0020700

by using the information possessed by a text from which data is to be extracted.

As a specific evaluation method, for example, the following methods are available.

(1) Evaluation of the degree of reliability according to an event type

An event type is extracted from a partial text, and the degree of reliability of the partial text is evaluated (Note that an event type is extracted from a partial text by using a database in which a target event, an attribute, and a partial text are corresponded).

(2) Evaluation of the degree of reliability according to the degree of attention

The degree of reliability is evaluated by noting the degree of attention of a target object within the text.

(3) Evaluation of the degree of reliability according to bibliographic information

The degree of reliability is evaluated according to bibliographic information (an author, publication media, etc.) possessed by a text. For example, if a text is a newspaper article, its degree of reliability is evaluated depending on whether the newspaper is either a popular paper or a quality paper as a news source.

0960760 "0960760

Here, an event type is attached in correspondence with a word included in an article. To be more specific, a database in which a correspondence between a particular word and an event type is made is generated. For example, an event type "personnel reshuffle" is attached to words such as "new president", "inauguration", etc., an event type "obituary notice" is attached to words such as "death", and the like.

Next, the data aggregating unit 12 calculates the degree of reliability as a data aggregation from individual degrees of reliability in order to aggregate data having the degree of reliability.

The following examples are considered as this algorithm.

i) The highest degree of reliability among the degrees of reliability of individual data is defined to be the degree of reliability of a data aggregation.

ii) An average of the degrees of reliability of individual data is defined to be the degree of reliability of a data aggregation.

The correctness/incorrectness determining unit 15 determines which data is correct based on the degree of reliability of a data aggregation and the numbers of occurrences. The following examples are considered as this algorithm.

0020070050960

i) The highest degree of reliability among the degrees of reliability of individual data is defined to be correct, and the remaining data are defined to be incorrect.

ii) A threshold value of the degree of reliability is set, and data having a particular or higher degree of reliability is defined to be correct.

Fig. 8 exemplifies a first internal configuration of the reliability assigning unit 16 shown in Fig. 7. This example shows the configuration in the case where the degree of reliability is evaluated according to an event type defined in the above described (1).

In Fig.8, 11 is the above described data extracting unit extracting object data from a text. The data extracting unit 11 analyzes a description of fact data within a text, and extracts the description as data as stated earlier. Assume that original texts are "a person B is inaugurated as a representative of a company A", "President D of a company A passed away", and "a company A puts B on the market" as shown in Fig. 9A. In this case, the "company A" is extracted as a target object, the "representative" and "product" are extracted as attribute names, and the "person B", "President D", and "B" are extracted as attribute values.

0052607.2865960

16 is the reliability degree assigning unit. An event type extracting unit 16a within the reliability degree assigning unit 16 extracts the keyword group such as the one shown in Fig. 9B from the original texts, and determines that a corresponding event type is possessed if a keyword included in the text matches any of the values within the table. As a result, event types are extracted from the partial texts being extraction targets shown in Fig. 10A.

The reliability evaluating unit 16b evaluates the degrees of reliability of fact data according to the event types by referencing an event type/reliability degree correspondence table 16d shown in Fig. 9D, as illustrated by Fig. 10B. If the degree of reliability of fact data which does not correspond to an event type defaults to, for example, 0.5.

By assigning the degree of reliability as described above, the degree of reliability can accurately be evaluated with the use of the knowledge such that, for example, the degree of reliability of an obituary notice is higher than that of an article regarding personnel reshuffle because especially careful checking is made to the personal data in an obituary notice.

Fig. 11 exemplifies a second internal

0956897.0956900

configuration of the reliability degree assigning unit. This example shows the configuration in the case where the degree of reliability is evaluated according to the degree of attention in the above described (2).

In Fig. 11, 11 is an object data extracting unit extracting object data itself, 16 is a reliability degree assigning unit, 16e is an attention degree evaluating unit evaluating the degree of attention of an object to be extracted within a text, and 16f is a reliability degree evaluating unit evaluating the degree of reliability according to the degree of attention.

As the method evaluating the degree of attention, which is executed by the attention degree evaluating unit 16e, the following algorithms can be considered.

Note that the explanation given below mainly refers to Japanese-language text, but it is easily recognized that similar algorithms could be applied to text in another language by one skilled in the art.

i) Examining a postpositional particle which immediately succeeds a target object, and the degree of attention of an object followed by a modifying postpositional particle such as "は", "も", etc. is defined to be the highest. The degree of attention is defined to be low in other cases. (Examining whether

0969807.092700

or not a target object is a subject word. If the target object is a subject word, the degree of attention is defined to be the highest. If not, and the degree of attention is defined to be low in other cases. For example, as shown in Fig. 12A, the degrees of attention of the subject word attached with the modifying postpositional particle, an object word, and the other element are respectively set to be 0.8, 0.5, and 0.4. It is determined whether or not the object data within an original text is either the subject or the object word, or the other element is determined as shown in Fig. 12B. Then, the degree of attention is set according to the determination result.

ii) The position of a target object within a text (the order of the target object from the beginning), or the order of the original sentence including the target object in a paragraph is counted, and the degree of attention of the target object word is evaluated by using a correspondence table between the position of a word and the degree of attention.

For example, the degree of attention is set according to the position of object data within an original text by using the correspondence table between the position of a word and the degree of attention as shown in Fig. 12C.

The reliability degree evaluating unit 16f calculates the degree of reliability to be possessed by fact data by using the degree of attention extracted as described above. Fundamentally, an evaluation algorithm is set so that the degree of reliability goes up as the degree of attention of a target object rises. By way of example, as shown in Fig. 12D, it is determined whether or not the degree of attention is higher than a threshold value α, and the degree of reliability is set according to a result of the determination. For the text analysis at this time, which word is a subject or an object word, etc. is obtained by using an existing program.

As described above, an accurate correctness/incorrectness determination can be made by raising the degree of reliability of an object to which attention is paid with the use of the information of a modifying postpositional particle, the position of a target object within a text, etc.

Fig. 13 shows a third internal configuration of the reliability degree assigning unit. This example show the configuration in the case where the degree of reliability is evaluated according to the bibliographical information in the above described (3).

In Fig. 13, 11 is the data extracting unit

0096960.7650250

extracting object data itself as described above, and 16 is the reliability degree assigning unit which receives as an input the bibliographical information (issuance source, author, etc.) possessed by a text, and examines the degree of reliability to be possessed by fact data with the use of a bibliographical information/reliability degree correspondence table 16h.

For example, the degree of reliability of a text is evaluated according to an issuance source, and a corresponding degree of reliability is assigned depending on whether or not the degree of reliability of the issuance source is high.

Hereinafter, explanation is provided with reference to the specific examples shown in Figs. 14A through 14E. Assume that bibliographical information (issuance sources) corresponding to the descriptions of original texts are respectively "news office A", "news office B", and "new agency C" as shown in Fig. 14A, and their degrees of reliability are respectively set to be 0.6, 0.8, and 0.9 in the bibliographical information/reliability degree correspondence table 16h as shown in Fig. 14B. In this case, the reliability degree assigning unit 16 assigns the degree of reliability to each of the texts according to the

bibliographical information/reliability degree correspondence table 16h, and the degree of reliability according to a news source is assigned to the object data output from the data extracting unit 11 as shown in Fig. 14C.

The above described data aggregating unit 12 shown in Fig. 3 aggregates the fact data assigned with the degree of reliability by the algorithm in the above described i) or ii), and passes the aggregated data to the inconsistency detecting unit 13. Since the representative of the company A is "B" and "D" among the fact data shown in Fig. 14C and an inconsistency exists, the inconsistency detecting unit 13 recognizes the representative of the company A "B" and "D" as inconsistent data, assigns the degrees of reliability, and outputs the fact data to the correctness/incorrectness determining unit 14 as shown in Fig. 14D.

The correctness/incorrectness determining unit 14 makes a correctness/incorrectness determination, for example, by using the algorithm in the above described i) or ii). By way of example, if the correctness/incorrectness determination is made by the algorithm in i) in which the data having the highest degree of reliability is selected to be correct and other

09659697.092700

data are recognized to be incorrect, the representative of the company A "B" is discarded as an error, and "D" is recognized to be correct and output to the data integrating unit 15. As a result, the data shown in Fig. 14F is output from the data integrating unit 15.

Fig. 15 exemplifies the configuration for generating the bibliographical information/reliability degree correspondence table 16h within the reliability degree assigning unit shown in Fig. 13.

In this figure, fact data to which a correctness/incorrectness flag is attached is input to a bibliographical information attribute scanning unit 17. The correctness/incorrectness flag indicates whether fact data is either correct or incorrect. This flag may be manually attached beforehand or automatically attached by a different system.

The bibliographical information attribute scanning unit 17 searches the whole of data for each attribute value of bibliographical information, etc., and extracts the fact data possessed by an attribute. Assume that the degrees of reliability of news sources such as the above described news office A, news office B, and news agency C are obtained. In this case, the whole of the data is searched for each of the news offices

and news agency, and fact data to which a correctness/incorrectness flag is attached is extracted.

A reliability degree evaluating unit 18 calculates the correct answer ratio of the data based on the correctness/incorrectness flag for the fact data extracted by the bibliographical information attribute scanning unit 17, and obtains the degree of reliability for each bibliographical information. As a result, the respective degrees of reliability of, for example, the above described news offices A and B and news agency C can be obtained.

A data registering unit 19 registers the degrees of reliability obtained by the reliability degree evaluating unit 18 to the bibliographical information/reliability degree correspondence table 16h, and puts the degrees into a database.

By generating the bibliographical information/reliability degree correspondence table 16h as described above, the operation for manually registering data to a correspondence table can be eliminated.

Fig. 16 shows a third preferred embodiment according to the present invention. This preferred embodiment is intended to decide the determination

0966897 0966897

method used by the correctness/incorrectness determining unit 14 by adding a determination method deciding unit 20 to the configuration shown in Figs. 4 and 7. The other constituent elements are the same as those shown in Figs. 4 and 7.

In Fig. 16, when an inconsistent data group is input to the correctness/incorrectness determining unit 14, the determination method deciding unit 20 first examines the target object and the attribute name of fact data. Then, the determination method deciding unit 20 references the bibliographical information/reliability degree correspondence table 21, and decides the method for a correctness/incorrectness determination. A target object, an attribute name, and a determination method according thereto are pre-registered to the the bibliographical information/reliability degree correspondence table correspondence table.

For example, if a plurality of persons corresponding to an attribute name such as a division director exist, a first determination method such as a method with which all of data having the degree of reliability that is equal to or higher than a threshold value is registered to the bibliographical information/reliability degree correspondence table

correspondence table 21. Or, if only one person corresponding to an attribute name such as "president" exists, a second determination method with which only the data having the highest degree of reliability is determined to be correct is registered to the above described table. The determination method deciding unit 20 specifies the first determination method if an attribute name is a division director, or specifies the second method if an attribute name is a president.

The correctness/incorrectness determining unit 14 determines whether each data within a data group is either correct or incorrect with the correctness/incorrectness determination method specified by the determination method deciding unit 20.

By making a correctness/incorrectness determination as described above, a unique correctness/incorrectness determination can be made for data which can possess a plurality of values, such as a division director of a company, and data only allowed to possess a unique value such as a president.

Fig. 17 shows a fourth preferred embodiment according to the present invention. This preferred embodiment is intended to discard data determined to be incorrect as single data by adding an error pattern removing unit 22 to the above described first or the

JP2000-076060

second preferred embodiment. The other constituent elements are the same as those shown in Fig. 4 or 7.

In Fig. 17, the error pattern removing unit 22 is arranged between the data aggregating unit 12 and the inconsistency detecting unit 13. The error pattern removing unit 22 discards data determined to be incorrect as single data by referencing an error pattern database 23, when the data is given from the data aggregating unit 12.

Figs. 18A-18C exemplify an error pattern determination made by the error pattern removing unit 22 shown in Fig. 17. These figures show an example where an error is detected by stipulating a telephone number that does not begin with "0" as an error pattern of a telephone number.

For example, if the data extracted by the data extracting unit 11 are the telephone numbers of companies A and B as shown in Fig. 18A, the error pattern removing unit 22 references the error pattern database 23, and makes a comparison between the above described telephone numbers and the error pattern of telephone numbers.

Here, assume that the error pattern shown in Fig. 18B is registered to the error pattern database 23. Fig. 18B describes that a telephone number that does not begin

0960907-0960907-0960907

with "0" is an error in a normal expression. Namely, the normal expression shown in Fig. 18B is a normal expression of a UNIX program for making a string matching, and "^" indicates the beginning, "[^0]" indicates a number which is not "0", and "[0-9]+" indicates a string of one or more numerals from 0 to 9. Here, the process for "-" is ignored.

Since the telephone number of the company B "119-0003" begins with a numeral other than "0", the error pattern removing unit 22 determines that this number is an error as a result of the comparison between the telephone numbers shown in Fig. 18A and the error pattern shown in Fig. 18B, and discards the telephone number of the company B.

Fig. 19 shows a fifth preferred embodiment according to the present invention. This preferred embodiment is intended to cope with representation fluctuations by arranging a data unifying unit to unify data having similar attribute values. The other constituent elements are the same as those shown in Fig. 7.

In Fig. 19, a data unifying unit 24 unifies data having similar attribute values by referencing a data fluctuations database 25. As a result, a correctness/incorrectness determination can be

EP 0 597 630 A2

prevented from being erroneously made as if not so many representation fluctuations occur for each representation, although many representation fluctuations occur actually.

Figs. 20A through 20C exemplifies the process performed in this preferred embodiment. When data shown in Fig. 20A is extracted by the data extracting unit 1, the data unifying unit 24 unifies data having similar attribute values.

A condition such that a name including a family name and a name having only a family name can be unified as similar data is assumed to be set as a unification condition of name data in the data fluctuations database 25 in this example. The data unifying unit 24 unifies the data having "Ichiro Yamada" as an attribute value of an attribute name "representative" of a company A and the data having "Yamada" under the above described condition by referencing the data fluctuations database 25. Consequently, the data of the attribute name "representative" of the company A are unified as shown in Fig. 20B, and the frequency of the data is set to be a total of the number of occurrences of both the data.

When the data are unified by the data unifying unit 24 as described above, a correctness/incorrectness determination is made according to the unified frequency

0027609.0986960

by the correctness/incorrectness determining unit 14. Suppose that the correctness/incorrectness determination is made with the above described algorithm in which data having the maximum number of occurrences within a group is determined to be correct, and the others are determined to be incorrect. In this case, as the "representative" of the company A, "Taro Yamada" is determined to be correct, while "Taro Suzuki" is determined to be incorrect as shown in Fig. 20C.

Because the number of occurrences of "Taro Suzuki" is larger than the respective numbers of occurrences of "Ichiro Yamada" and "Yamada" in this example, "Taro Suzuki" is determined to be correct if data unification is not performed. However, a proper correctness/incorrectness can be made by performing the above described data unification.

As stated earlier, the following effects can be obtained according to the present invention.

(1) Fact data is extracted from a text, data of the same type among the extracted data are unified, a data aggregation is made throughout the text, an inconsistent data group which cannot be consistent is detected by scanning an aggregated data set, which data is correct is determined within the inconsistent data group, and correct fact data are unified by removing error data,

0096997.0022500

whereby suitable data can be integrated by removing an error portion from the errors and fluctuations within extracted data, which are caused by an erroneous description within a text or an extraction process error.

(2) The degree of reliability is assigned to fact data when the data is extracted from a text, and whether each data within a data group is either correct or incorrect is determined by using the degree of reliability, whereby the accuracy of the correctness/incorrectness determination can be improved.

(3) A determination method used when a correctness/incorrectness determination is made is specified according to an attribute name, and the correctness/incorrectness determination is made by the specified determination method, whereby a flexible correctness/incorrectness determination can be made according to an attribute.

(4) A matching between extracted fact data and a pre-registered error pattern is made, and the extracted fact data is determined to be incorrect and is discarded when a match is found between the extracted data and the pre-registered error pattern, whereby it becomes possible to remove an error that can be determined alone.

(5) Similar data are unified, and inconsistency

0996977.0007500

detection is made after the similar data are unified into one, whereby fluctuations caused by different expression of the same thing can be absorbed.

0096897.0092700

What is claimed is:

1. A fact data unifying method, comprising:

extracting from a text fact data stipulated by a combination of a target object, an attribute name, and an attribute value;

grouping data of a same type among extracted fact data, and performing a data aggregation throughout a text;

detecting an inconsistent data group which cannot be consistent by scanning an aggregated data set; and

determining which data is correct within the inconsistent data group, and unifying correct fact data by removing incorrect data.

2. A fact data unifying apparatus, comprising:

a data extracting unit extracting from a text fact data stipulated by a combination of a target object, an attribute name, and an attribute value;

a data aggregating unit grouping data of a same type among fact data extracted by said data extracting unit, and aggregating the number of occurrences throughout a text;

an inconsistency detecting unit detecting an inconsistent data group which cannot be consistent by

scanning a data set aggregated by said data aggregating unit;

a correctness/incorrectness determining unit determining which data is correct within the inconsistent data group detected by said inconsistency detecting unit; and

a final data integrating unit integrating correct data aggregated by said data aggregating unit, and data determined to be correct by said correctness/incorrectness determining unit.

3. The fact data unifying apparatus according to claim 2, further comprising

a reliability degree assigning unit assigning a degree of reliability to fact data when the fact data is extracted from a text, wherein:

the degree of reliability of aggregated data is calculated from the degrees of reliability of individual data, and assigned to an aggregation result, when the numbers of occurrences are aggregated by said data aggregating unit; and

said correctness/incorrectness determining unit determines whether each data within a data group is either correct or incorrect by using the degree of reliability assigned to the data.

JP2007-286960

4. The fact data unifying apparatus according to claim 3, wherein:

said reliability degree assigning unit comprises

an event type extracting unit determining a type of event information possessed by a text from which fact data is to be extracted when the fact data is extracted from a text, and

a reliability degree evaluating unit evaluating the degree of reliability according to an event type based on a correspondence table between an event type and the degree of reliability.

5. The fact data unifying apparatus according to claim 3, wherein:

said reliability degree assigning unit comprises

an attention degree evaluating unit calculating a degree of attention to a target object to be extracted within a text, and

a reliability degree evaluating unit evaluating the degree of reliability of data based on the degree of attention.

6. The fact data unifying apparatus according to claim 3, wherein

said reliability degree assigning unit comprises:

a bibliographical information/ reliability degree correspondence table making a correspondence between bibliographical information of an issuance source, an author of a text, etc., and the degree of reliability of each data described in the text; and

a reliability degree evaluating unit evaluating the degree of reliability of a text according to bibliographical information of a text by referencing said bibliographical information/reliability degree correspondence table, when data is extracted from the text.

7. The fact data unifying apparatus according to claim 6, wherein

said bibliographical information/reliability degree correspondence table is generated by attaching a correctness/incorrectness flag to fact data extracted by said data extracting unit, by receiving as an input the fact data to which the correctness/incorrectness flag is attached, and by calculating an expectation value of correctness/incorrectness of data having a particular attribute value for each attribute name of the fact data.

8. The fact data unifying apparatus according to claim 2, further comprising:

an attribute/determination method correspondence table which makes a correspondence between a target object, an attribute name, and a determination method used when a correctness/incorrectness determination is made; and

a determination method deciding unit deciding a correctness/incorrectness determining method according to an attribute based on said attribute/determination method correspondence table, wherein

said correctness/incorrectness determining unit makes a correctness/incorrectness determination by a method specified by said determination method deciding unit, when an inconsistent data group is input.

9. The fact data unifying apparatus, wherein:

an error pattern removing unit is arranged between said data extracting unit and said inconsistency detecting unit; and

said error pattern removing unit makes a correctness/incorrectness determination for each data by making a matching between the fact data extracted by said data extracting unit and a pre-registered error

JP2002269960

pattern, determines and discards the extracted fact data as an error if the extracted fact data matches the pre-registered error pattern, and transmits only data determined to be correct to said inconsistency detecting unit.

10. The fact data unifying apparatus according to claim 2, wherein:

a data integrating unit arranged after said data aggregating unit; and

said data integrating unit passes integrated data to said inconsistency detecting unit after integrating similar data into one.

11. A storage medium on which is recorded a program for causing an information processing device to execute a process for unifying fact data stipulated by a combination of a target object, an attribute name, and an attribute value, which are extracted from a text, said process comprising:

extracting from a text fact data stipulated by a combination of a target object, an attribute name, and an attribute value;

grouping data of a same type among extracted fact data, and performing a data aggregation throughout a

JP2002-269960

text;

detecting an inconsistent data group which cannot be consistent by scanning an aggregated data set; and

determining which data is correct within the inconsistent data group, and unifying correct fact data by removing incorrect data.

DOCID: <4980950>

Abstract of the Disclosure

A data extracting unit extracts from a text fact data stipulated by a combination of a target object, an attribute name, and an attribute value. A data aggregating unit groups data similar to the extracted fact data throughout the text, and aggregates the number of occurrences. An inconsistency detecting unit detects an inconsistent data group which cannot be consistent by scanning a data set aggregated by the data aggregating unit. A correctness/incorrectness determining unit determines which data is correct within the inconsistent data group. A final data integrating unit integrates and outputs correct data. Additionally, the degree of reliability is assigned to fact data when the fact data is extracted from a text, and also a correctness/incorrectness determination of each data within a data group can be made by using the degree of reliability assigned to the fact data.

09559887.062700

0076076 "7698960

FIG. 1A

EXAMPLE OF CORRESPONDENCE TABLE

[EXPRESSION FORMAT] *2 IS ASSIGNED AS PRESIDENT OF *1

FIG. 1B

EXTRACTED DATA

TARGET OBJECT VALUE	ATTRIBUTE NAME	ATTRIBUTE
*1	REPRESENTATIVE	*2

FIG. 1C

PROCESS EXAMPLE

[INPUT TEXT]
PERSON D (MATCHING *2) IS ASSIGNED AS NEW PRESIDENT OF JOINT COMPANY C (MATCHING *1) ESTABLISHED BY COMPANIES A AND B

FIG. 1D

EXTRACTED DATA

TARGET OBJECT VALUE	ATTRIBUTE NAME	ATTRIBUTE
COMPANY C	REPRESENTATIVE	PERSON D

PRIOR ART

007260"26869960

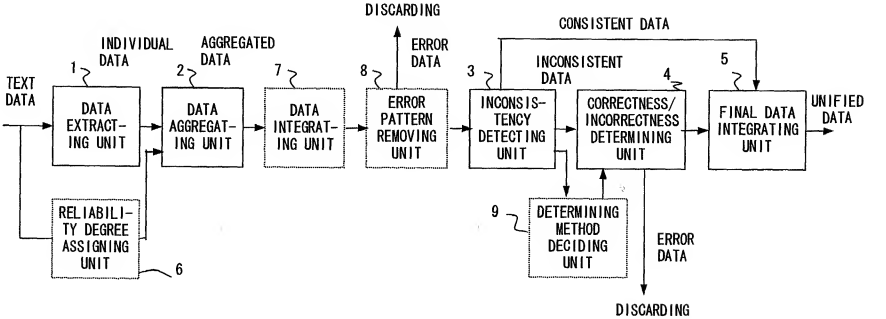


FIG. 2

0966987 "092700

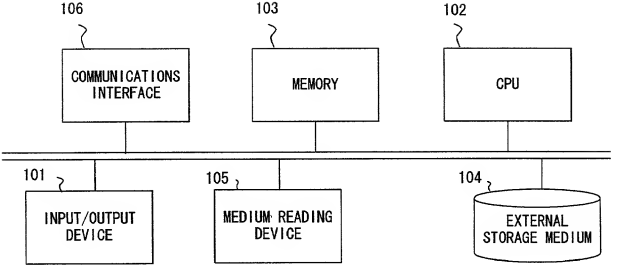


FIG. 3

096869897"092700

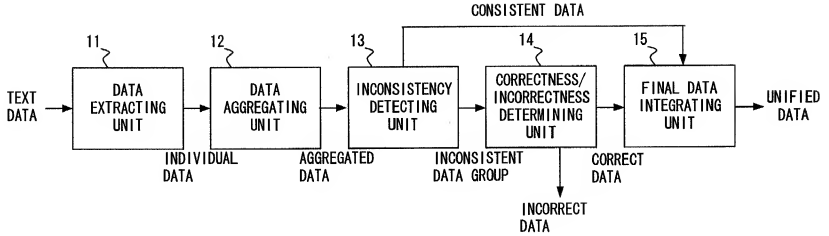


FIG. 4

FIG. 5A

OUTPUT EXAMPLE OF DATA EXTRACTING UNIT

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE
COMPANY A	REPRESENTATIVE	B
COMPANY F	REPRESENTATIVE	G
COMPANY A	REPRESENTATIVE	B
COMPANY A	REPRESENTATIVE	D
COMPANY H	LOCATION	COUNTRY C

FIG. 5B

OUTPUT EXAMPLE OF DATA AGGREGATING

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE	THE NUMBER OF OCCURRENCES
COMPANY A	REPRESENTATIVE	B	2
COMPANY F	REPRESENTATIVE	G	1
COMPANY A	REPRESENTATIVE	D	1
COMPANY H	LOCATION	COUNTRY C	1

FIG. 5C

OUTPUT EXAMPLE OF INCONSISSTENCY DETECTING UNIT

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE	THE NUMBER OF OCCURRENCES
COMPANY A	REPRESENTATIVE	B	2
COMPANY A	REPRESENTATIVE	D	1

FIG. 5D

OUTPUT EXAMPLE OF CORRECTNESS/INCORRECTNESS DETERMINING UNIT

DATA			CORRECTNESS/INCORRECTNESS DETERMINATION
COMPANY A	REPRESENTATIVE	B	CORRECT
COMPANY A	REPRESENTATIVE	D	INCORRECT →DISCARDING

FIG. 5E

OUTPUT EXAMPLE OF FINAL DATA INTEGRATING UNIT

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE
COMPANY A	REPRESENTATIVE	B
COMPANY F	REPRESENTATIVE	G
COMPANY H	LOCATION	COUNTRY C

0966997-092700

09/250,760 02/26/99

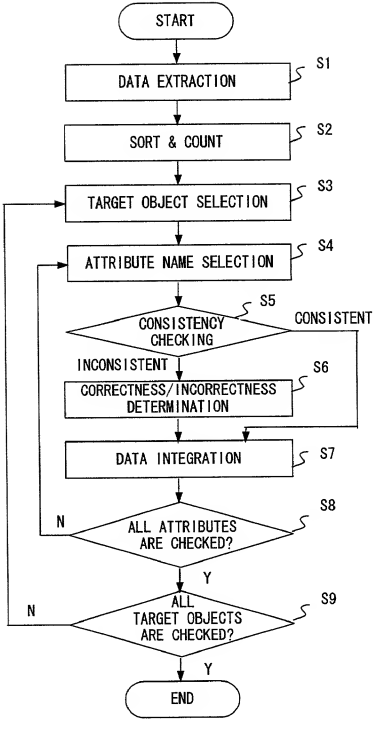
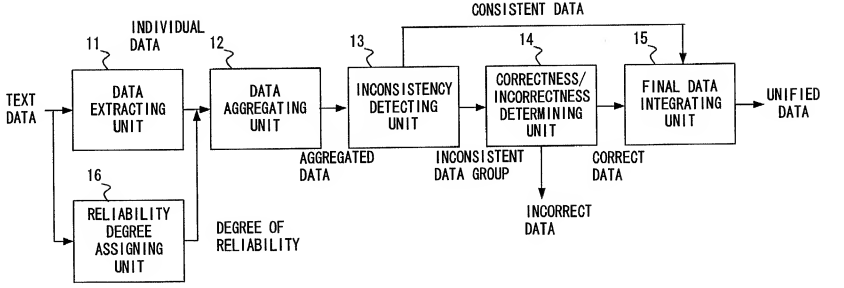


FIG. 6

09689897 .092700



F I G. 7

0960897.092700

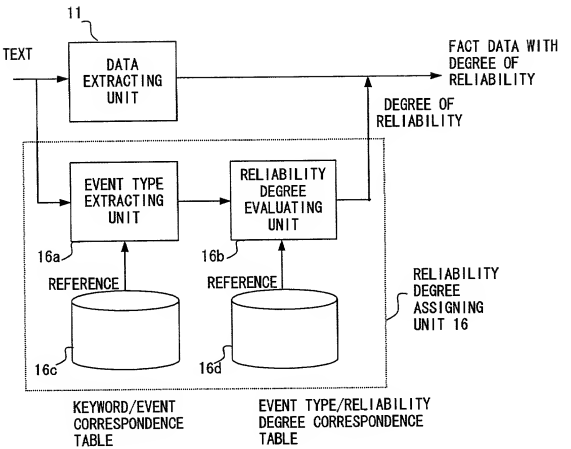


FIG. 8

EXAMPLE OF ORIGINAL TEXT AND EXTRACTED DATA

ORIGINAL TEXT	TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE
PERSON B IS INAUGURATED AS REPRESENTATIVE OF COMPANY A	COMPANY A	REPRESENTATIVE	B
PRESIDENT D OF COMPANY A PASSED AWAY	COMPANY F	REPRESENTATIVE	PRESIDENT D
COMPANY A PUTS B ON THE MARKET	COMPANY A	PRODUCT	B

FIG. 9A

EXAMPLE OF CORRESPONDENCE TABLE BETWEEN ORIGINAL TEXT AND KEY WORD

ORIGINAL TEXT	EXTRACTED KEYWORDS
PERSON B IS INAUGURATED AS REPRESENTATIVE OF COMPANY A	COMPANY A, COMPANY A, PERSON B, INAUGURATED
PRESIDENT D OF COMPANY A PASSED AWAY	COMPANY A, PRESIDENT D, PASS AWAY
COMPANY A PUTS B ON THE MARKET	COMPANY A, B, PUT ON THE MARKET

FIG. 9B

EXAMPLE OF KEYWORD/EVENT TYPE CORRESPONDENCE TABLE

KEYWORD	EVENT TYPE	DEGREE OF RELIABILITY
INAUGURATED, DISMISS	PERSONNEL RESHUFFLE	0.8
PASS AWAY	OBITUARY NOTICE	0.9

FIG. 9C

EXAMPLE OF EVENT TYPE/DEGREE OF RELIABILITY CORRESPONDENCE TABLE

EVENT TYPE	DEGREE OF RELIABILITY
PERSONNEL RESHUFFLE	0.8
OBITUARY NOTICE	0.9
DEFAULT	0.5

FIG. 9D

0022607-960930

002260'26869960

DETERMINATION EXAMPLE OF EVENT TYPE OF TEXT

ORIGINAL TEXT	EXTRACTED KEYWORDS	EVENT TYPE
PERSON B IS INAUGURATED AS REPRESENTATIVE OF COMPANY A	COMPANY A, REPRESENTATIVE, PERSON B, INAUGURATED	PERSONNEL RESHUFFLE
PRESIDENT D OF COMPANY A PASSED AWAY	COMPANY A, PRESIDENT B, PASS AWAY	OBITUARY
COMPANY A PUTS B ON THE MARKET	COMPANY A, B, PUT ON THE MARKET	DEFAULT

FIG. 10A

EXAMPLE OF DEGREE OF RELIABILITY ASSIGNED TO EACH DATA

ORIGINAL TEXT	EVENT TYPE	DEGREE OF RELIABILITY
PERSON B IS INAUGURATED AS REPRESENTATIVE OF COMPANY A	PERSONNEL RESHUFFLE	0.8
PRESIDENT D OF COMPANY A PASSED AWAY	OBITUARY NOTICE	0.9
COMPANY A PUTS B ON THE MARKET	DEFAULT	0.5

FIG. 10B

09668697.092700

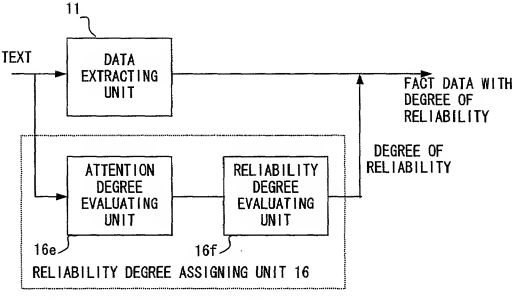


FIG. 11

EXAMPLE OF METHOD EVALUATING DEGREE OF ATTENTION

FIG. 12A

SUBJECT WORD	0.8
OBJECT WORD	0.5
OTHER ELEMENTS	0.4

EXAMPLE OF DEGREE OF ATTENTION ASSIGNED TO OBJECT WITHIN ORIGINAL TEXT (THE DEGREE OF ATTENTION IS SET BY RECOGNIZING THE DEGREES OF ATTENTION OF SUBJECT AND OBJECT WORDS TO BE HIGHER IN THIS ORDER)

FIG. 12B

ORIGINAL TEXT	A社の	B社長は	新製品群を発表
	⇩	⇩	⇩
DEGREE OF ATTENTION	0.4	0.8	0.5

ORIGINAL TEXT	A大臣は	B社長と	懇談
	⇩	⇩	
DEGREE OF ATTENTION	0.8	0.4	

FIG. 12C EXAMPLE OF CORRESPONDENCE TABLE BETWEEN WORD POSITION AND DEGREE OF ATTENTION

POSITION $<$ 5	DEGREE OF ATTENTION = 5 − POSITION
POSITION= $>$ 5	DEGREE OF ATTENTION = 0

FIG. 12D EXAMPLE OF ALGORITHM EVALUATING DEGREE OF RELIABILITY

DEGREE OF ATTENTION $>$ α	DEGREE OF RELIABILITY= 0.9
DEGREE OF ATTENTION $\leqq$ α	DEGREE OF RELIABILITY= 0.7

0096997.0093700

0669897-052700

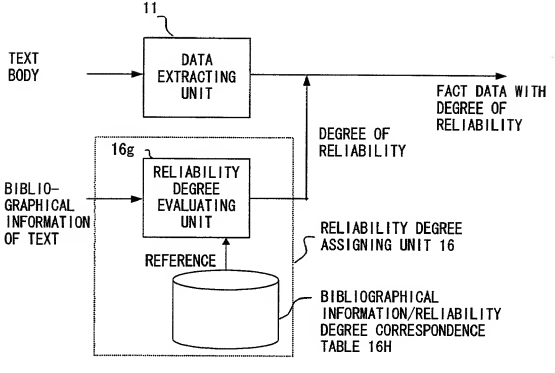


FIG. 13

FIG. 14A EXAMPLE OF DESCRIPTION WITHIN ORIGINAL TEXT AND BIBLIOGRAPHICAL INFORMATION

TEXT	MEDIA
PRESIDENT B OF COMPANY A	A NEWS OFFICE
COMPANY A (REPRESENTATIVE D)	B NEWS OFFICE
COMPANY A (HEADQUARTERS: E-SHI, C PREFECTURE)	C NEWS AGENCY

FIG. 14B EXAMPLE OF BIBLIOGRAPHICAL INFORMATION/RELIABILITY DEGREE CORRESPONDENCE TABLE

MEDIA	DEGREE OF RELIABILITY
A NEWS OFFICE	0.6
B NEWS OFFICE	0.8
C NEWS AGENCY	0.9

FIG. 14C EXAMPLE OF OUTPUT OF DATA INTEGRATING UNIT

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE	DEGREE OF RELIABILITY
COMPANY A	REPRESENTATIVE	B	0.6
COMPANY A	REPRESENTATIVE	D	0.8
COMPANY H	LOCATION	COUNTRY C	0.9

FIG. 14D EXAMPLE OF OUTPUT OF INCONSISTENCY DETECTING UNIT

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE	DEGREE OF RELIABILITY
COMPANY A	REPRESENTATIVE	B	0.6
COMPANY A	REPRESENTATIVE	D	0.8

FIG. 14E EXAMPLE OF DETERMINATION MADE BY CORRECTNESS/INCORRECTNESS DETERMINING UNIT

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE	DEGREE OF RELIABILITY	CORRECT/ INCORRECT
COMPANY A	REPRESENTATIVE	B	0.6	INCORRECT
COMPANY A	REPRESENTATIVE	D	0.8	CORRECT

FIG. 14F EXAMPLE OF OUTPUT OF INCONSISTENCY DETECTING UNIT

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE
COMPANY A	REPRESENTATIVE	B

0096897-0092700

09689897.092700

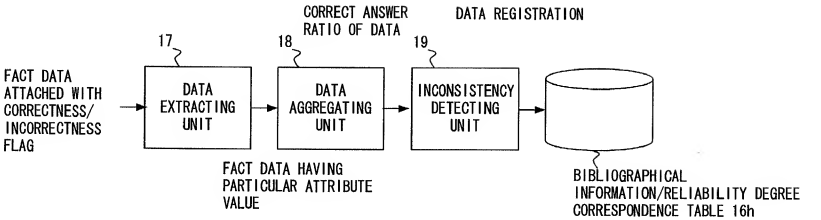


FIG. 15

0025-7685990

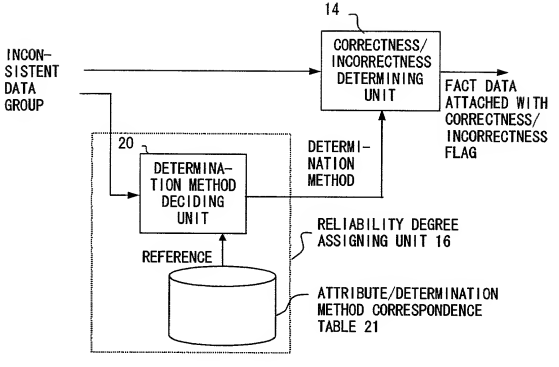


FIG. 16

007260"27869897

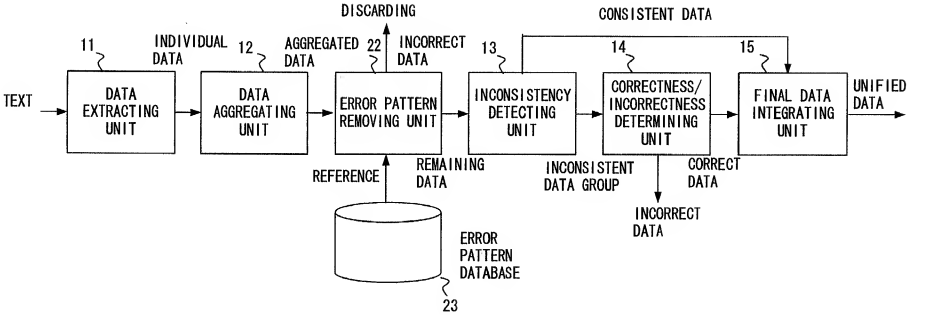


FIG. 17

EXAMPLE OF EXTRACTED DATA

COMPANY A	TELEPHONE NUMBER	03-356-7098
COMPANY B	TELEPHONE NUMBER	119-0003

FIG. 18A

EXAMPLE OF ERROR PATTERN

ATTRIBUTE NAME	NORMAL EXPRESSION	MEANING
TELEPHONE NUMBER	^[^0] [0-9]+	NUMBER THAT DOES NOT BEGIN WITH "0"

FIG18B

EXAMPLE OF CORRECTNESS/INCORRECTNESS DETERMINATION

DATA			CORRECTNESS/ INCORRECTNESS
COMPANY A	TELEPHONE NUMBER	03-356-7098	CORRECTNESS
COMPANY B	TELEPHONE NUMBER	119-0003	INCORRECTNESS

FIG. 18C

0960997.092700

007260"26869960

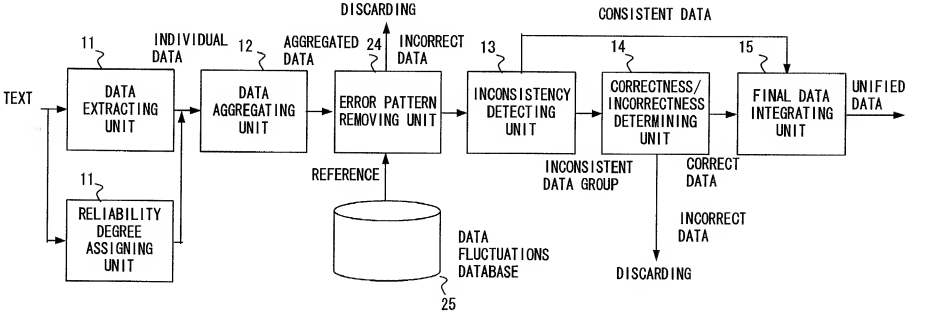


FIG. 19

EXAMPLE OF EXTRACTED DATA

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE	FREQUENCY
COMPANY A	REPRESENTATIVE	ICHIRO YAMADA	20
COMPANY A	REPRESENTATIVE	YAMADA	30
COMPANY A	REPRESENTATIVE	TARO SUZUKI	30

FIG. 20A

EXAMPLE OF DATA UNIFYING PROCESS

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE	FREQUENCY
COMPANY A	REPRESENTATIVE	ICHIRO YAMADA (20 OCCURRENCES OF YAMADA ARE ADDED)	40
COMPANY A	REPRESENTATIVE	TARO SUZUKI	30

FIG. 20B

EXAMPLE OF CORRECTNESS/INCORRECTNESS DETERMINATION

TARGET OBJECT	ATTRIBUTE NAME	ATTRIBUTE VALUE	FREQUENCY	CORRECTNESS/ INCORRECTNESS
COMPANY A	REPRESENTATIVE	ICHIRO YAMADA	40	CORRECT
COMPANY A	REPRESENTATIVE	TARO SUZUKI	30	INCORRECT

FIG. 20C

0096697-090700